

Discriminating Intended Human Objects in Consumer Videos

Hiroshi Uegaki, Yuta Nakashima and Noboru Babaguchi
 Division of Electrical, Electronic and Information Engineering
 Graduate School of Engineering, Osaka University
 2-1 Yamadaoka, Suita, Osaka, 565-0871, Japan
 {uegaki, nakashima, babaguchi}@nanase.comm.eng.osaka-u.ac.jp

Abstract—In a consumer video, there are not only intended objects, which are intentionally captured by the camcorder user, but also unintended objects, which are accidentally framed-in. Since the intended objects are essential to present what the camcorder user wants to express in the video, discriminating the intended objects from the unintended objects are beneficial for many applications, e.g., video summarization, privacy protection, and so forth. In this paper, focusing on human objects, we propose a method for discriminating the intended human objects from the unintended human objects. We evaluated the proposed method using 10 videos captured by 3 camcorder users. The results demonstrate that the proposed method successfully discriminates the intended human objects with 0.45 of recall and 0.80 of precision.

Keywords-intended object; intended human object;

I. INTRODUCTION

In a consumer video, not only intended objects, which are intentionally captured by the camcorder user, but also unintended objects, which are accidentally framed-in, are captured. Since the intended objects play an important role to present what he/she wants to express in the video, discriminating the intended objects from the unintended objects is beneficial in many applications including video summarization, video adaptation, privacy protection, and so forth.

Visual attention models have been used for these applications. Ma et al. proposed a method for video summarization, consisting of visual, audio, and linguistic attention models [1]. Wang et al. also proposed a method for generating a static video summary using a visual attention model [2]. In these methods, the visual attention models are used to extract key-frames, video skims, and regions of interest in the frames. Fan et al. and Liu et al. proposed methods to adapt a video to small displays [3], [4]. These methods determine which regions are cropped from the frames based on visual saliency. Hua et al. presented automated home video editing system [5]. Their system utilizes a visual attention model to select suitable video segments from a home video. Itti evaluated the applicability of a visual attention model for video compression, and suggested that the model was useful to improve the compression ratio [6]. Most of the visual attention models, however, used in these applications only provide whether each pixel is visually salient or not, and this may spoil what camcorder user wants to express when the



Figure 1. An example of intended and unintended human objects. The red circle and the black rectangle indicate the intended human objects and the unintended human objects, respectively.

intended objects are not visually salient. Nevertheless, to the best of our knowledge, this problem has not been resolved so far.

In this paper, to tackle this problem, we propose a method for discriminating the intended human objects from the unintended human objects considering that a human is one of the most important objects. Figure 1 shows an example of intended and unintended human objects. The camcorder follows the intended human object surrounded by the red circle, while the unintended human object surrounded by the black rectangle that appears in (a) is framed out in (b). The camcorder user moves the camcorder so that the intended human objects are placed at arbitrary positions in the frames or are followed. However, he/she does not care much for the unintended human objects. Therefore, we consider that the camcorder user's intention is reflected to the features related to how he/she moves the camcorder in accordance with each human object. In addition, we consider that the camcorder user's intention rarely changes in a short interval. Based on these discussions, we construct camcorder user's intention model, and the proposed method discriminates whether or not a human object is intended frame-by-frame using the model.

The rest of the paper is organized as follows: In Section II, we describe the proposed method and Section III presents experimental results. We conclude the paper in Section IV.

II. DISCRIMINATING INTENDED HUMAN OBJECTS

Based on the discussions in the previous section, the camcorder user's intention is modeled by the hidden Markov

model (HMM). The proposed method first tracks each human object in a video and extracts features. Then, using the camcorder user's intention model, the proposed method discriminates whether or not the human object is intended frame-by-frame.

A. Feature extraction

In the proposed method, each of the human objects in the video is tracked to extract features and to apply the HMM. First, we detect a human object by a skin color detector assuming that a human is captured including his/her face. Although some methods for face detection which detect frontal or profile faces were proposed such as [7], they are computationally expensive. The skin color detector enables us to detect human object with a low computational cost.

Then, each of the detected skin color regions is tracked using the continually adaptive mean-shift (CAMSHIFT) [8] and features related to how the camcorder user moves the camcorder, i.e., area $S_{n,i}$, centroid position $(g_{n,i}^x, g_{n,i}^y)$, and motion ratio $F_{n,i}$, are extracted from the i -th tracked skin color region r_i in the n -th frame f_n . The motion ratio is defined as the ratio of the amplitude of the human object's motion to that of the camera motion among the successive two frames. The smaller the motion ratio is, the more likely the camcorder follows the human object.

The procedure to detect and initialize a CAMSHIFT tracker is as follows:

- 1) Convert f_n into the HSV color space.
- 2) Construct a binary image which represents whether or not the pixel is skin color using empirically determined thresholds T_1 and T_2 for the hue value as

$$b_{n,i} = \begin{cases} 1 & \text{if } T_1 \leq h_{n,i} \leq T_2 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $b_{n,i}$ and $h_{n,i}$ denote the i -th pixel value in the binary image and the hue value in f_n , respectively. The pixel value 1 in the binary image stands for the skin color.

- 3) Remove noises in the binary image using a morphological closing. The morphological closing is the dilation followed by the erosion.
- 4) Label each of the connected components in the binary image. Each label represents the skin color region r'_j .
- 5) Initialize a CAMSHIFT tracker with a search window that is centered at the centroid of r'_j and is of size $M \times M$ if r'_j is not tracked. That is, if the following condition is satisfied for all i , a CAMSHIFT tracker is initialized:

$$|\gamma_{n,j}^x - g_{n,i}^x| \geq T_G \quad \text{or} \quad |\gamma_{n,j}^y - g_{n,i}^y| \geq T_G \quad (2)$$

where $(\gamma_{n,j}^x, \gamma_{n,j}^y)$ and $(g_{n,i}^x, g_{n,i}^y)$ are the centroid of r'_j and that of r_i , which is the region already tracked by CAMSHIFT tracker, in the n -th frame. T_G is a predetermined threshold.

The CAMSHIFT tracker is terminated when the size of the tracked skin color region becomes zero.

The area, $S_{n,i}$, and the centroid position, $(g_{n,i}^x, g_{n,i}^y)$ of r_i are computed from the search window of the corresponding CAMSHIFT tracker. The motion ratio $F_{n,i}$ is calculated as follows:

$$F_{n,i} = \frac{\sqrt{(g_{n,i}^x - g_{n-1,i}^x)^2 + (g_{n,i}^y - g_{n-1,i}^y)^2}}{\sqrt{(m_n^x)^2 + (m_n^y)^2} + 1} \quad (3)$$

where (m_n^x, m_n^y) denotes the translation of the frames from f_{n-1} to f_n . The translation is obtained by modeling the camera motion as the projective transform and using [9]. We define a feature vector sequence as

$$X_i = \{\mathbf{x}_{n,i} | n = \alpha_i, \dots, \beta_i\} \quad (4)$$

where $\mathbf{x}_{n,i} = (S_{n,i}, g_{n,i}^x, g_{n,i}^y, F_{n,i})$ and the tracking of the skin color region r_i is started from α_i -th frame and is ended at β_i -th frame.

B. Camcorder user's intention model

$y_{n,i} = 0$, and 1 denote r_i in f_n is unintended and intended, respectively. We model $\mathbf{x}_{n,i}$ as a random variable that follows the Gaussian mixture model (GMM) given $y_{n,i}$. That is,

$$p(\mathbf{x}_{n,i} | y_{n,i}) = \sum_{k=1}^K w_{y_{n,i}}^k \mathcal{N}(\mathbf{x}_{n,i} | \mu_{y_{n,i}}^k, \Sigma_{y_{n,i}}^k) \quad (5)$$

where K , $w_{y_{n,i}}^k$, $\mu_{y_{n,i}}^k$, and $\Sigma_{y_{n,i}}^k$ are the number of mixture components, the weight, the mean and the covariance of the k -th mixture component for $y_{n,i}$, respectively, and $\mathcal{N}(\cdot | \mu, \Sigma)$ is the multivariate normal distribution with the mean and covariance of μ and Σ , respectively. The parameters of the GMM are estimated using the EM algorithm.

As aforementioned, the camcorder user's intention does not change frequently. To take this into account, we employ the HMM. The hidden states of the HMM is $y_{n,i}$ and its transition is described by the transition matrix T , i.e.,

$$T = \begin{pmatrix} t_{0,0} & t_{0,1} \\ t_{1,0} & t_{1,1} \end{pmatrix} \quad (6)$$

where $t_{u,v}$ ($u, v \in \{0, 1\}$) stands for the transition from u to v . The transition matrix is obtained by counting the transitions. The initial probabilities, $\pi = (\pi_0, \pi_1)$, is determined empirically.

Let θ denote the parameters of the HMM. The joint distribution of X_i and $\mathbf{y}_i = (y_{\alpha_i,i}, y_{\alpha_i+1,i}, \dots, y_{\beta_i,i})$ is represented as

$$p(X_i, \mathbf{y}_i | \theta) = \pi_{y_{\alpha_i,i}} \prod_{n=\alpha_i+1}^{\beta_i} t_{y_{n-1,i}, y_{n,i}} p(\mathbf{x}_{n,i} | y_{n,i}). \quad (7)$$

Table I
THE PARAMETERS.

T_1	T_2	T_G	M	K	π_0	π_1
0	30	200	10	3	0.5	0.5

The proposed method discriminates whether or not a human object is intended by maximizing $P(X_i, \mathbf{y}_i | \theta)$ using the Viterbi algorithm with respect to \mathbf{y}_i , i.e.,

$$\hat{\mathbf{y}}_i = \arg \max_{\mathbf{y}_i} p(X_i, \mathbf{y}_i | \theta). \quad (8)$$

III. EXPERIMENTS

A. Experimental setup

We evaluated the performance of the proposed method using the leave-one-out cross-validation. The dataset used in the evaluation contains 10 videos captured by 3 camcorder users, each of which is of size 720×480 and 30 frames per second. The dataset includes videos capturing the persons playing card game, playing catch, eating dishes, and so forth. The face regions in these videos were manually specified, and the camcorder users annotated whether each of the face region was intended or not as a ground truth. The parameters of the GMM and the transition matrix were determined using only correctly tracked skin color regions. The rest of the parameters were empirically determined as listed in Table I.

We evaluated the following two cases; (1) the performance of the discrimination by the camcorder user's intention model only, and (2) the performance of the proposed method. The performance in (2) includes the errors of the human detection and CAMSHIFT tracking, i.e., skin color regions which were not detected/tracked and were not correspond to actual faces may be judged to be incorrect discriminations. Thus, we need to evaluate (1) to obtain the performance of camcorder user's intention model itself.

To evaluate the performance of (1), we calculated Recall_d and Precision_d defined as

$$\text{Recall}_d = \frac{\text{TPN}}{\text{TPN} + \text{FNN}}, \quad \text{Precision}_d = \frac{\text{TPN}}{\text{TPN} + \text{FPN}}$$

where TPN, FNN, and FPN are the true positive number, the false negative number, and the false positive number, respectively. These indices were calculated only from the tracked skin color regions which made the correspondences with the manually specified face regions.

To evaluate the performance of (2), we computed Recall_w and Precision_w defined as

$$\text{Recall}_w = \frac{\text{TPN}}{\text{SPE}}, \quad \text{Precision}_w = \frac{\text{TPN}}{\text{DIS}}.$$

SPE and DIS stand for the total number of intended face regions specified by camcorder users and the total number of skin color regions which were discriminated to be intended by the model, respectively.

B. Experimental results

Table II shows the Recall_d and Precision_d . The videos such as VIDEO5 and VIDEO8 contained almost only one intended human objects, resulting in the large value of Precision_d . The reason why the value of Recall_d averaged over the 10 videos was low is because the intended human objects in the dataset were concentrated around the center of the frame, and thus, the many intended human objects which were not around the center of the frames were discriminated incorrectly as shown in Figure 2 (a). In addition, the camcorder users often did not move the camcorder when the motions of the intended human objects were small. This resulted in incorrect discrimination even for the case where only one intended human object was around the center the frame. Features related to trajectories of the human objects can alleviate this problem. However, even when the camcorder panned or zoomed to follow a human objects, the motion ratio enabled the model to discriminate correctly.

Table III shows the Recall_w and Precision_w . The camcorder user's intention model incorrectly discriminated for some skin color regions, which did not correspond to actual face regions, as the intended human objects. Thus, the values of Precision_w were lower than Precision_d . Recall_w were lower than Recall_d because some intended human objects were not detected or tracked. Example frames of VIDEO1 in Figure 2 (b) show that the table in the frames was detected and tracked as a skin color region, and was discriminated as an intended human object. In addition, some faces like those of the persons in both sides of the right most frame were not detected or tracked. Therefore, we need to improve the performances of human detection and tracking by employing another method.

IV. CONCLUSION

In this paper, we proposed a method for discriminating intended human objects from unintended human objects in consumer videos. Focusing on that the camcorder user's intention is reflected to how the camcorder is moved by the camcorder user and that the camcorder user's intention is not change frequently, we modeled the camcorder user's intention using the HMM. An experimental evaluation demonstrated that the proposed method discriminated the intended human objects with 0.45 of recall and 0.80 of precision including the performance of the human detection and tracking, while the performance of the camcorder user's intention model were 0.55 of recall and 0.93 of precision. This result shows that our method depends much on the

Table II
EVALUATION RESULTS FOR (1).

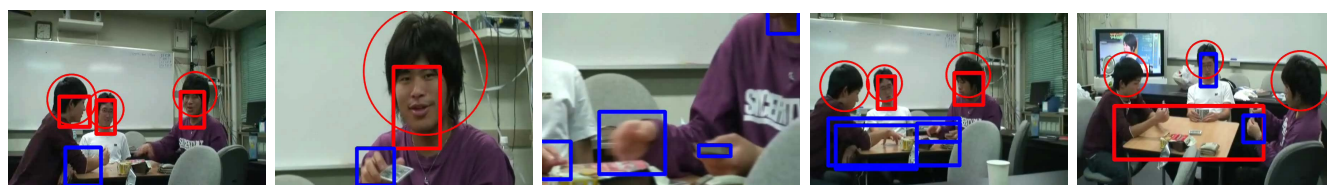
	VIDEO1	VIDEO2	VIDEO3	VIDEO4	VIDEO5	VIDEO6	VIDEO7	VIDEO8	VIDEO9	VIDEO10	AVE.
Recall _d @	0.44	0.57	0.73	0.50	0.74	0.70	0.42	0.37	0.35	0.68	0.55
Precision _d	0.90	0.99	0.97	0.95	1.00	0.58	1.00	1.00	1.00	0.91	0.93

Table III
EVALUATION RESULTS FOR (2).

	VIDEO1	VIDEO2	VIDEO3	VIDEO4	VIDEO5	VIDEO6	VIDEO7	VIDEO8	VIDEO9	VIDEO10	AVE.
Recall _w	0.33	0.55	0.64	0.49	0.73	0.61	0.23	0.24	0.33	0.36	0.45
Precision _w	0.57	0.70	0.97	0.94	1.00	0.53	0.49	1.00	1.00	0.83	0.80



(a) Example frames of VIDEO9



(b) Example frames of VIDEO1

Figure 2. Example frames of two videos. The circles and rectangles denote the face regions which are specified by the camcorder user and the skin color regions tracked by the CAMSHIFT trackers, respectively. The intended and unintended human objects are surrounded by the red and blue lines, respectively. First low (a) shows the intended human objects which were not around the center of the frames were sometimes discriminated incorrectly. Second low (b) shows the table in the frames was detected and tracked as a skin color region and was discriminated as an intended human objects, and some faces were not detected or tracked.

performance of the human detection and tracking. Our future work includes employing another human detector and adopting other features. This work is partly supported by a Grant-in-Aid for scientific research from the Japan Society for the Promotion of Science.

REFERENCES

- [1] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. A user attention model for video summarization. *In Proc. 10th ACM Intl. Conf. Multimedia*, pp. 533–542, 2002.
- [2] Tang Wang, Tao Mei, Xian-Sheng Hua, Xue-Liang Liu, and He-Qin Zhou. Video collage: a novel presentation of video sequence. *In Proc. IEEE ICME*, pp. 1479–1482, Jul. 2007.
- [3] Xin Fan, Xing Xie, He-Qin Zhou, and Wei-Ying Ma. Looking into video frames on small displays. *In Proc. 11th ACM Intl. Conf. Multimedia*, pp. 247–250, 2003.
- [4] Feng Liu and Michael Gleicher. Video retargeting: automating pan and scan. *In Proc. 14th ACM Intl. Conf. Multimedia*, pp. 241–250, 2006.
- [5] Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. Optimization-based automated home video editing system. *IEEE Trans. Circuits and Systems for Video Technology*, vol. 14, no. 1, pp.572–583, May 2004.
- [6] Laurent Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Trans. Image Processing*, vol. 13, no. 10, pp.1304–1318, Oct. 2004.
- [7] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. *In Proc. IEEE CVPR*, vol. 1, pp. 511–518, Dec. 2001.
- [8] Gary R. Bradski. Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal*, 1998.
- [9] Frederic Dufaux and Janusz Konrad. Efficient, robust, and fast global motion estimation for video coding. *IEEE Trans. Image Processing*, vol. 9, no. 3, pp. 497–501, Mar. 2000.